DeepLINK: Deep learning inference using knockoffs with applications to genomics

Zifan Zhu[a], Yingying Fan[b,1], Yinfei Kong[c], Jinchi Lv[b], and Fengzhu Sun[a,1]

[a]Quantitative and Computational Biology Department, University of Southern California, Los Angeles, CA 90089; [b]Data Sciences and Operations Department, Marshall School of Business, University of Southern California, Los Angeles, CA 90089; and [c]Department of Information Systems and Decision Sciences, California State University, Fullerton, CA 92831

Edited by Michael S. Waterman, University of Southern California, Los Angeles, CA, and approved July 16, 2021 (received for review March 10, 2021)

We propose a deep learning–based knockoffs inference framework, DeepLINK, that guarantees the false discovery rate (FDR) control in high-dimensional settings. DeepLINK is applicable to a broad class of covariate distributions described by the possibly nonlinear latent factor models. It consists of two major parts: an autoencoder network for the knockoff variable construction and a multilayer perceptron network for feature selection with the FDR control. The empirical performance of DeepLINK is investigated through extensive simulation studies, where it is shown to achieve FDR control in feature selection with both high selection power and high prediction accuracy. We also apply DeepLINK to three real data applications to demonstrate its practical utility.

false discovery rate | knockoffs | deep learning | microbiome | single-cell

The era of big data gives us enormous new opportunities but meanwhile, also produces unprecedented challenges in solving various data-related problems. The challenges are not just because of the large size of the data but also and even more caused by the complexity in, for example, text, image, video, and audio data. As a result, complicated models such as deep neural networks have been proposed and popularly used to analyze big data. Despite the appealing high prediction and classification power of deep neural networks, there is strong pushback from the scientific community because of its “black box” nature. The complicated structure of many deep neural networks has made the interpretation and reproducibility of such models incredibly difficult if even possible at all. To alleviate these issues, dimension reduction methods such as variable selection and latent factor models have been used in statistics and related applications.

In the past decade, feature (variable) selection has been a central topic in statistics (1, 2). Feature selection aims at identifying the truly important features that contribute to the effect of some response of interest. One desirable property of feature selection methods is that the error rate of selecting incorrect features can be controlled at some preselected target level while achieving high power. The celebrated procedure of Benjamini and Hochberg (3, 4) for false discovery rate (FDR) control has been shown to enjoy such a property both theoretically and empirically under some conditions of the P values calculated for evaluating the feature importance. Although a vast number of methods have been proposed for feature selection with the goal of controlling error rate, such as the Benjamini–Yekutieli procedure (5), local FDR (6), q value (7), the adaptive Benjamini–Hochberg procedure (BH for short hereafter) (8), P value weighting (9), FDR regression (10), independent hypothesis weighting (11), adaptive shrinkage (12), adaptive P value thresholding (13), and the structure-adaptive Benjamini–Hochberg algorithm (14), very few can be used in complicated models such as deep neural networks. The intrinsic difficulties are that most existing methods were proposed under much simpler model settings that are difficult or not even possible at all to generalize or depend heavily on the P values as the feature importance measure. Such P values can be calculated based on some classical or asymptotic theory in simpler models. When we move away from these simple model settings to more complicated ones such as deep neural networks, however, we no longer have the luxury of calculating theoretically justified P values, making feature selection highly challenging. Recently, Candès et al. (15) proposed a new framework of model-X knockoffs for achieving the FDR control in feature selection, bypassing the use of conventional P values. Model-X knockoffs can be used as a wrapper by combining with any feature selection methods that produce feature importance measures satisfying certain conditions. We provide a brief review of the model-X knockoffs in a later section. Thanks to the flexibility of model-X knockoffs, it was recently extended to the setting of deep neural networks in ref. 16 via proposing a new network architecture, DeepPINK, when the features have joint Gaussian distributions. The distributional assumption of joint Gaussian limits the practical applicability of the proposed method therein. In this paper, we explore more general distributional assumptions for the feature vector and propose a method for deep learning inference using knockoffs, named DeepLINK.

Latent factor models, which use lower-dimensional unobservable factors to model the comovements of features, have been well studied and broadly used in statistics (17–19), sociology (20, 21), bioinformatics (22–24), and economics (25–29). The most commonly used factor model assumes a linear relationship between the feature vector and latent factors. Since in practice, we can never be certain whether the dependency is truly linear, we are likely to face the problem of model

Significance

Although practically attractive with high prediction and classification power, complicated learning methods often lack interpretability and reproducibility, limiting their scientific usage. A useful remedy is to select truly important variables contributing to the response of interest. We develop a method for deep learning inference using knockoffs, DeepLINK, to achieve the goal of variable selection with controlled error rate in deep learning models. We show that DeepLINK can also have high power in variable selection with a broad class of model designs. We then apply DeepLINK to three real datasets and produce statistical inference results with both reproducibility and biological meanings, demonstrating its promising usage to a broad range of scientific applications.

Author contributions: Y.F., J.L., and F.S. designed research; Z.Z., Y.F., Y.K., J.L., and F.S. performed research; Z.Z., Y.F., Y.K., J.L., and F.S. analyzed data; and Z.Z., Y.F., Y.K., J.L., and F.S. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: fanyingy@usc.edu or fsun@usc.edu.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2104683118/-/DCSupplemental.

Published September 3, 2021.

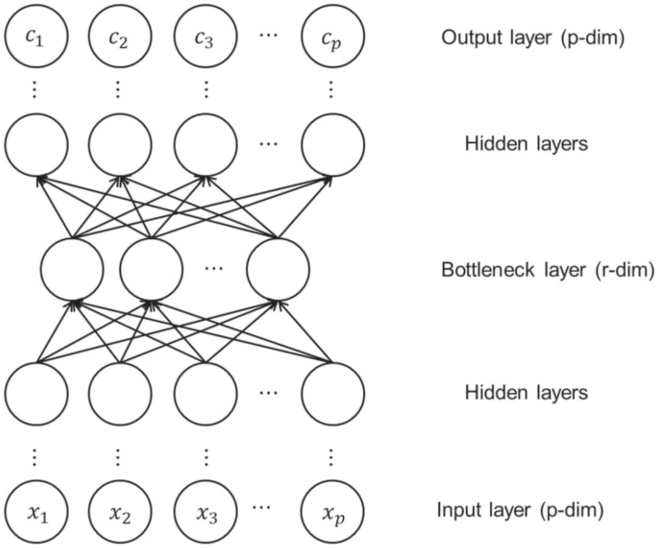

Fig. 1. The autoencoder architecture. p-dim and r-dim indicate p dimensions and r dimensions, respectively.

misspecification, making the statistical estimation and inference results unreliable. Recently, the advent of deep learning has motivated the nonlinear factor models described by the architecture of the autoencoder.

DeepLINK combines the flexible nonlinear factor modeling power of the autoencoder with the feature selection and prediction power of DeepPINK. The nonlinear factor model for the feature vector described by the autoencoder enables us to generate the knockoff variables effectively without imposing restrictive joint distribution assumptions (e.g., Gaussian) on features. The feature selection and prediction power of DeepPINK allow for interpretable and reproducible statistical inference without sacrificing much power. It is worth mentioning that for the special case when both the factor model and the regression model of response on features are linear, the problem of model-X knockoffs inference was investigated in ref. 30 via proposing a parametric inference framework of Intertwined Probabilistic Factors Decoupling (IPAD). We demonstrate the superior performance of DeepLINK via simulations and three real data examples. Compared with IPAD, DeepLINK is more flexible and more robust to model misspecification and meanwhile, achieves comparable feature selection results with generally higher power.

Deep Learning–Based Knockoffs Inference

Variable Selection with False Discovery Rate (FDR) Control. Consider the high-dimensional supervised learning with independent and identically distributed (i.i.d.) observations $(\mathbf{x}_i, y_i)$, $i = 1, \ldots, n$, where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T \in \mathbb{R}^p$ is the feature vector and $y_i \in \mathbb{R}$ is the scalar response. The number of features p can be comparable with or even larger than the number of observations n. Let $\{1, 2, \cdots, p\}$ be the full set of all the features. Assume that the conditional distribution of response y_i depends only on a small subset of features, and we aim to find the Markov blanket (31) (i.e., the smallest subset $\mathcal{S}_0$ such that y_i is independent of all remaining features given those in $\mathcal{S}_0$). That is,

$$y_i \perp\!\!\!\perp \{x_{ij} : j \in \mathcal{S}_0^c\} \mid \{x_{ik} : k \in \mathcal{S}_0\}, \tag{1}$$

where $\mathcal{S}_0^c$ denotes the complement of subset $\mathcal{S}_0$ in the full set $\{1, 2, \cdots, p\}$. The existence and uniqueness of the Markov blanket can be guaranteed under mild conditions on the joint distribution of $(\mathbf{x}_i, y_i)$. The discussions in ref. 15 have more details. For the ease of presentation, we refer to features in $\mathcal{S}_0$ as the "true" features and those in $\mathcal{S}_0^c$ as the "null" features in future presentation.

The goal of our study is to identify true features while controlling the error rate under a predetermined level. Various performance metrics have been proposed to measure the feature selection error rate, such as the familywise error rate, k-familywise error rate (k-FWER) (32), false discovery proportion (FDP) (33), and FDR (3). Here, we adopt the widely used FDR defined as

$$\mathrm{FDR} := \mathbb{E}[\mathrm{FDP}] \;\; \text{with FDP} := \frac{\left|\hat{\mathcal{S}} \cap \mathcal{S}_0^c\right|}{\max\{|\hat{\mathcal{S}}|, 1\}}, \tag{2}$$

where $\hat{\mathcal{S}}$ is the set of selected features using some statistical procedure, $|\cdot|$ means the cardinality of a set, and the expectation is taken with respect to the randomness in $\hat{\mathcal{S}}$. A modified version of FDR (mFDR) is defined as

$$\mathrm{mFDR} := \mathbb{E}\left[\frac{\left|\hat{\mathcal{S}} \cap \mathcal{S}_0^c\right|}{|\hat{\mathcal{S}}| + q^{-1}}\right], \tag{3}$$

where $q \in (0, 1)$ is the target FDR level. It is seen that FDR is more conservative than mFDR since controlling the FDR naturally results in the control of mFDR. We also use another important performance measure, power, to investigate the capability of a statistical procedure in discovering the true features. Formally speaking, power is defined as the expectation of the true discovery proportion (TDP):

$$\mathrm{Power} := \mathbb{E}[\mathrm{TDP}] \;\; \text{with TDP} := \frac{\left|\hat{\mathcal{S}} \cap \mathcal{S}_0\right|}{|\mathcal{S}_0|}. \tag{4}$$

A desirable inference framework should be able to control the FDR at a prechosen target level and meanwhile, achieve high power.

Model Settings. We focus on the setting where the high-dimensional feature vector $\mathbf{x}_i$ depends on some low-dimensional latent factor vector $\mathbf{f}_i \in \mathbb{R}^r$ with $r \ll p$ in a potentially nonlinear fashion. Specifically, assume the following factor structure for $\mathbf{x}_i$:

$$\mathbf{x}_i = \mathbf{g}(\mathbf{f}_i) + \boldsymbol{\epsilon}_i, \;\; i = 1, \cdots, n, \tag{5}$$

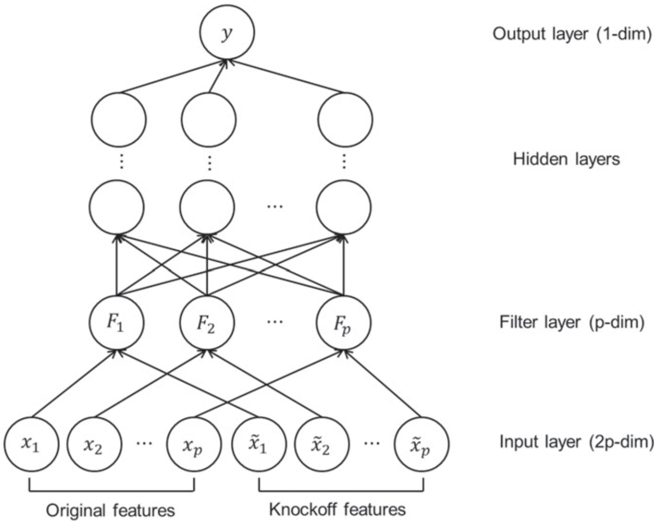

Fig. 2. The DeepPINK architecture. 2p-dim, p-dim, and 1-dim indicate 2p dimensions, p dimensions, and 1 dimension, respectively.

Table 1. Neural network parameter settings

	Activation	Loss	Optimizer	Regularization
Autoencoder				
Linear *h*	ELU	MSE	Adam	None
Nonlinear *h*	ELU	MSE	Adam	None
MLP				
Linear *h*	ELU	MSE	Adam	L_1 regularization
Nonlinear *h*	ELU	MSLE	Adam	L_1 regularization

where $\mathbf{g}$ is a vector-valued function whose coordinates can take some nonlinear functional forms that are unknown to us, and $\boldsymbol{\epsilon}_i \in \mathbb{R}^p$ is the factor model error vector with i.i.d. components. We make the additional assumption that the marginal distribution of the components of $\boldsymbol{\epsilon}_\mathbf{i}$ is from some parametric family $f_{\boldsymbol{\theta}}$ with unknown parameter $\boldsymbol{\theta} \in \mathbb{R}^m$, where m is some fixed positive integer.

When the coordinates of $\mathbf{g}$ are all linear functions, model Eq. **5** becomes the widely used latent factor model in the literature, which we will refer to as the linear factor model to ease the presentation. Most existing works have been developed under the linear factor model assumption, which can be restrictive in some applications. Our proposed method will use a data-adaptive way to estimate the possibly nonlinear function $\mathbf{g}$.

We also assume that the response y_i depends on $\mathbf{x}_i$ via the following nonparametric regression model

$$y_i = h(\mathbf{x}_i) + \varepsilon_i,\ i = 1, \cdots, n, \tag{6}$$

where h is some unknown function and can be either linear or nonlinear and ε_i's are independent model errors. For the ease of presentation, we will use matrix and vector notation by denoting $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$ the $n \times p$ design matrix, $\mathbf{F} = (\mathbf{f}_1, \ldots, \mathbf{f}_n)^T$ the $n \times r$ matrix of factors, and $\mathbf{y} = (y_1, \ldots, y_n)^T$ the n-dimensional response vector. Define $\mathbf{C}$ as an $n \times p$ matrix whose rows are $\mathbf{g}(\mathbf{f}_i)^T$, $i = 1, \ldots, n$. Then, model Eq. **5** can be rewritten as

$$\mathbf{X} = \mathbf{C} + \mathbf{E}, \tag{7}$$

where $\mathbf{E}$ is a matrix with the ith row being ϵ_i^T. Our goal can be specifically stated as developing a feature selection method with the FDR controlled at the target level q under the flexible model settings Eqs. **5** and **6**.

The Model-X Knockoffs Framework. We will adopt the recently developed model-X knockoffs framework introduced in ref. 15 to achieve our goal of feature selection. For completeness, we give a brief review of the model-X knockoffs framework below. We refer the readers to ref. 15 for full details.

As discussed in the Introduction, various FDR control methods have been proposed since the seminal work of BH. Most of these existing methods achieve the FDR control under the assumption that valid *P* values can be calculated. However, having valid *P* values can become a luxury in the high-dimensional big data settings. Taking the generalized linear models as an example, when the feature dimensionality p diverges with sample size n at a rate of $n^{2/3}$ or faster, the classical asymptotic theory of maximum likelihood estimation (MLE) no longer applies. Consequently, the resulting *P* values calculated using the formula from the classical asymptotic theory become invalid. Ref. 34 has

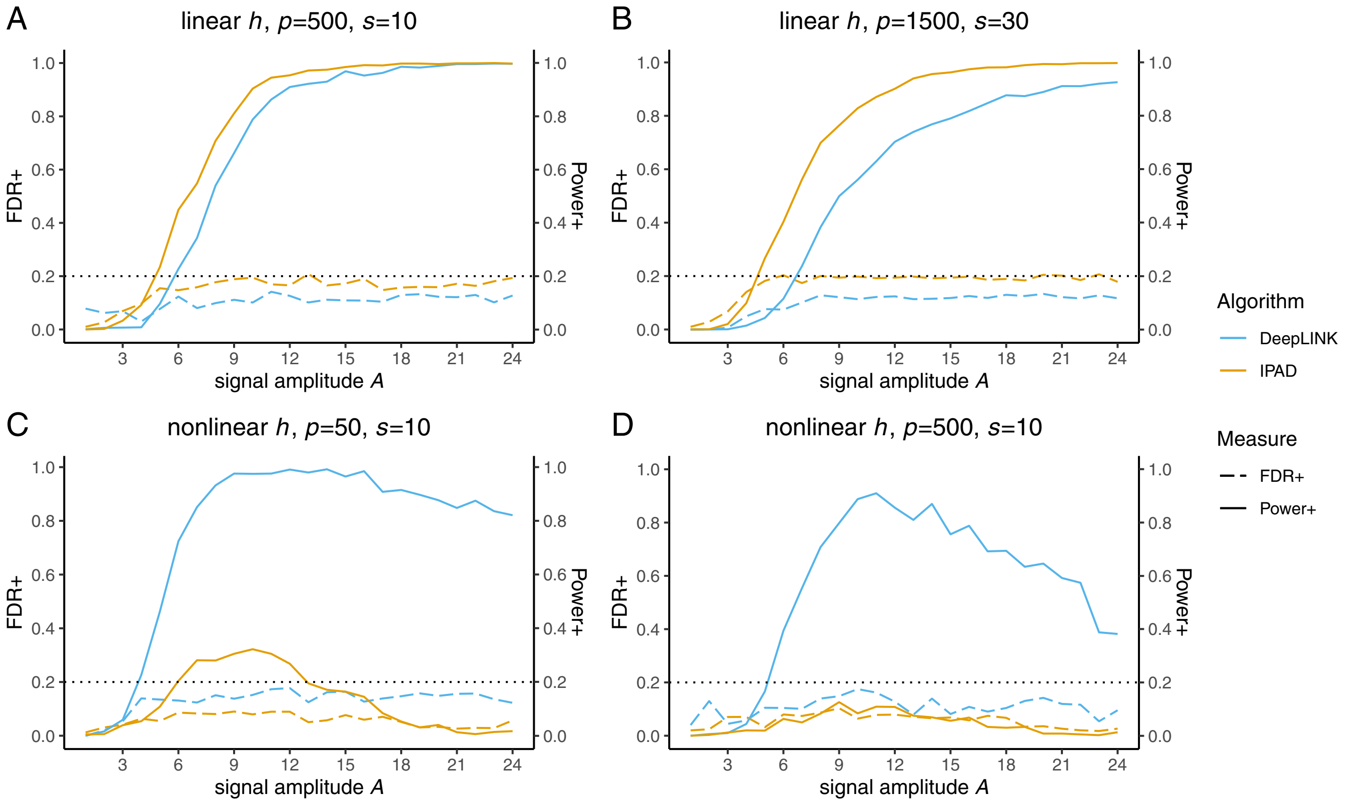

Fig. 3. Comparisons between DeepLINK and IPAD in simulation settings with the linear factor model. *A*–*D* represent different combinations of the link function *h*, the number of features *p*, and the number of true signals *s*. FDR+ and Power+ denote the empirical FDR and power obtained using the knockoff+ threshold. The black dashed lines indicate the target FDR level. Each plot shows the change of FDR+ (solid lines) and Power+ (long dashed line) for DeepLINK (blue) and IPAD (orange) against varying signal amplitude *A*.

formal results on such a phenomenon. When more complicated models such as the random forests or deep neural networks are used, how to calculate valid P values for evaluating the feature importance is still an open question. To overcome this difficulty, Barber and Candès (35) introduced the fixed-X knockoffs framework, bypassing the use of p values to achieve the FDR control in the Gaussian linear model when p is smaller than $n/2$. Recently, Candès et al. (15) proposed the model-X knockoffs framework, which achieves theoretically guaranteed FDR control in arbitrary dimensions and for arbitrary dependence structure of response y on features $\mathbf{x}$. These advantages motivate us to adapt the model-X knockoffs framework to our model settings.

The salient idea of the model-X knockoffs is to construct the so-called "model-X knockoff variables," which perfectly mimic the dependence structure of the original variables but are conditionally independent of the response. For completeness, we include the definition of the model-X knockoff variables introduced in ref. 15 as follows.

***Definition*:** Model-X knockoff variables for a set of random variables $\mathbf{x} = (x_1, \cdots, x_p)^T$ are a new set of random variables $\tilde{\mathbf{x}} = (\tilde{x}_1, \cdots, \tilde{x}_p)^T$ that satisfies the following properties.

1) For any subset $S \subset \{1, \cdots, p\}$, $(\mathbf{x}^T, \tilde{\mathbf{x}}^T)_{\text{swap}(S)} \overset{\mathcal{D}}{=} (\mathbf{x}^T, \tilde{\mathbf{x}}^T)$, where $(\mathbf{x}^T, \tilde{\mathbf{x}}^T)_{\text{swap}(S)}$ is obtained by swapping the components x_j and $\tilde{x}_j$ in $(\mathbf{x}^T, \tilde{\mathbf{x}}^T)$ for each $j \in S$ and $\overset{\mathcal{D}}{=}$ denotes equal in distribution;
2) $\tilde{\mathbf{x}} \perp\!\!\!\perp y \mid \mathbf{x}$.

The second property above is satisfied as long as $\tilde{\mathbf{x}}$ is constructed without using the information of response y. To construct knockoff variables that satisfy the first property, we need to know the joint distribution of $\mathbf{x}$. When such distribution is available, ref. 15 proposed a generic algorithm Sequential Conditional Independent Pairs (SCIP) for the knockoff variable construction. When such information is unavailable, there has been some recent work on the practical construction of knockoff variables (for example, refs. 30 and 36–40).

Denote by $\tilde{\mathbf{x}}_i$ the vector of knockoff variables for $\mathbf{x}_i$, $i = 1, \ldots, n$, and let $\tilde{\mathbf{X}} = (\tilde{\mathbf{x}}_1, \cdots, \tilde{\mathbf{x}}_n)^T$. For each $j = 1, \ldots, p$, let W_j be the knockoff statistic defined for measuring the importance of the jth original feature. Specifically, W_j is a function of the augmented data matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$ and the response vector $\mathbf{y}$ [i.e., $W_j = w_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y})$, with w_j a function satisfying the "sign-flip" property]:

$$w_j([\mathbf{X}, \tilde{\mathbf{X}}]_{\text{swap}(S)}, \mathbf{y}) = \begin{cases} w_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}), & j \notin S, \\ -w_j([\mathbf{X}, \tilde{\mathbf{X}}], \mathbf{y}), & j \in S, \end{cases} \qquad \textbf{[8]}$$

where S can be any subset of $\{1, \cdots, p\}$. The formal characterizations of the desired knockoff statistics as well as examples are in ref. 15. Intuitively, valid knockoff statistics measure the importance of original features, with large positive ones indicating the original features being important, and for unimportant features in $\mathcal{S}^c$, the corresponding W_j's are expected to have small magnitudes and be symmetric around zero.

Finally, the set of important features is selected as $\hat{\mathcal{S}} = \{j : W_j \geq t\}$ with $t = T$ or $t = T_+$, where T is the knockoff threshold and T_+ is the knockoff+ threshold as proposed in ref. 15 and included below for completeness:

$$T = \min\left\{ t > 0 : \frac{|\{j : W_j \leqslant -t\}|}{\max\{|\{j : W_j \geqslant t\}|, 1\}} \leqslant q \right\}, \qquad \textbf{[9]}$$

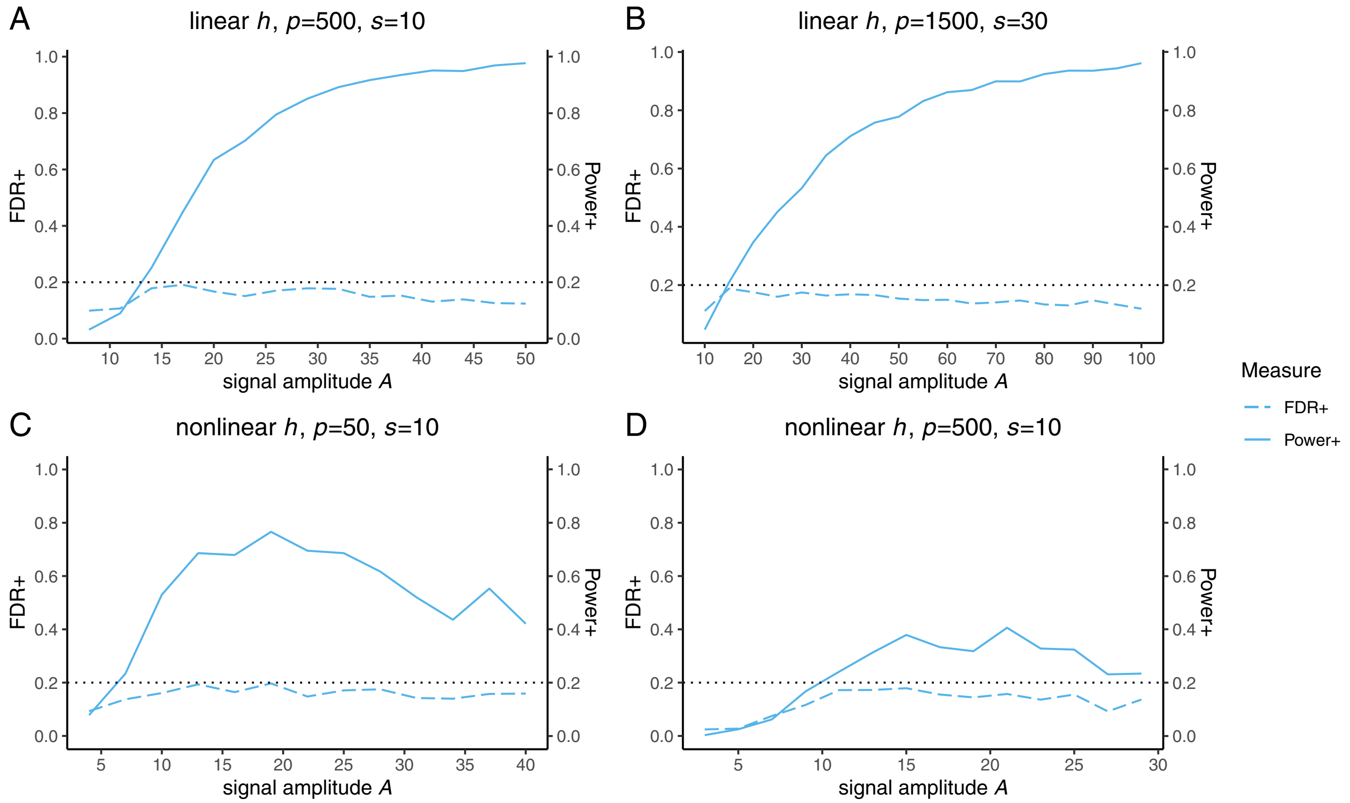

Fig. 4. Simulation results of DeepLINK in settings with the additive quadratic factor model. *A*–*D* represent different combinations of the link function *h*, the number of features *p*, and the number of true signals *s*. FDR+ and Power+ denote the empirical FDR and power obtained using the knockoff+ threshold. The black dashed lines indicate the target FDR level. Each plot shows the change of FDR+ (solid line) and Power+ (long dashed line) against varying signal amplitude *A*.

$$T_+ = \min\left\{t>0: \frac{1+|\{j: W_j \leqslant -t\}|}{\max\{|\{j: W_j \geqslant t\}|, 1\}} \leqslant q\right\}. \tag{10}$$

Here, $\hat{S}$ is defined as an empty set if $T=\infty$ or $T_+=\infty$.

It has been formally shown in ref. 15 that the knockoff threshold controls the mFDR exactly and the knockoff+ threshold controls the FDR exactly at the finite-sample level, regardless of the sample size n, feature dimensionality p, and dependence structure of response y on features $\mathbf{x}$.

DeepLINK. We next introduce our framework of DeepLINK, a deep learning–based statistical inference framework using knockoffs. It consists of two parts: 1) an autoencoder network for the knockoff variable construction and 2) a multilayer perceptron (MLP) network for feature selection with the FDR control.

As reviewed in the last section, there are two key ingredients in the successful implementation of the model-X knockoffs framework: 1) the construction of knockoff variables and 2) the construction of knockoff statistics. Since the joint distribution of $\mathbf{x}_i$ is unknown to us, the generic algorithm proposed in ref. 15 is no longer applicable to our settings. A remedy is to exploit the nonlinear factor model structure in Eq. **5** to construct approximate knockoff variables using the estimated distribution.

In view of Eq. **7**, ideally if the realization $\mathbf{C}$ and the marginal distribution $f_{\boldsymbol{\theta}}$ of ϵ_i are both known a priori, then the knockoff variables can be constructed as

$$\tilde{\mathbf{X}} = \mathbf{C} + \tilde{\mathbf{E}}, \tag{11}$$

with the entries of $\tilde{\mathbf{E}}$ independently drawn from distribution $f_{\boldsymbol{\theta}}$. It can be easily checked that such $\tilde{\mathbf{X}}$ satisfies the two properties in *Definition*. Since $\mathbf{C}$ and $f_{\boldsymbol{\theta}}$ are generally unknown in practice, we next discuss methods to estimate them. We will also discuss the construction of knockoff statistics.

Part 1: Autoencoder for knockoffs construction. The principal component analysis (PCA) has been a predominant method for extracting latent factors in the existing literature (41, 42). However, a key assumption for PCA to work well is that $\mathbf{x}_i$ depends on $\mathbf{f}_i$ linearly. To address the challenge caused by the potentially nonlinear factor model as specified in Eq. **5**, we propose to use the deep learning model of autoencoder.

Given the design matrix $\mathbf{X}$, we train an autoencoder with $\mathbf{X}$ as the input as well as the target output. An illustrative plot of the autoencoder network is shown in Fig. 1. Denote by $\hat{\mathbf{C}}$ the corresponding autoencoder output matrix. We propose to construct the knockoffs data matrix as

$$\tilde{\mathbf{X}} = \hat{\mathbf{C}} + \tilde{\mathbf{E}}, \tag{12}$$

where $\tilde{\mathbf{E}}$ is a matrix with entries independently sampled from the estimated marginal distribution $f_{\hat{\boldsymbol{\theta}}}$ of ϵ_i. For the specific case when $f_{\boldsymbol{\theta}}$ is the Gaussian density of $N(0,\sigma^2)$, we have $\boldsymbol{\theta}=\sigma^2$, which can be estimated as $\hat{\sigma}^2=(np)^{-1}\sum_{1\le i\le n, 1\le j\le p}\hat{e}_{ij}^2$ with $\hat{e}_{ij}$'s the entries of the residual matrix $\hat{\mathbf{E}}=\mathbf{X}-\hat{\mathbf{C}}$. This corresponds to the maximum likelihood estimate with the pseudo-observations $\hat{\mathbf{E}}$. In general, parameter $\boldsymbol{\theta}$ can be estimated by the maximum likelihood approach or the method of moments based on $\hat{\mathbf{E}}$.

Part 2: MLP for feature selection. To construct the knockoff statistics, we need to first construct the feature importance measure. Since an important goal of our framework is to accommodate the flexible nonlinear relationship between response y

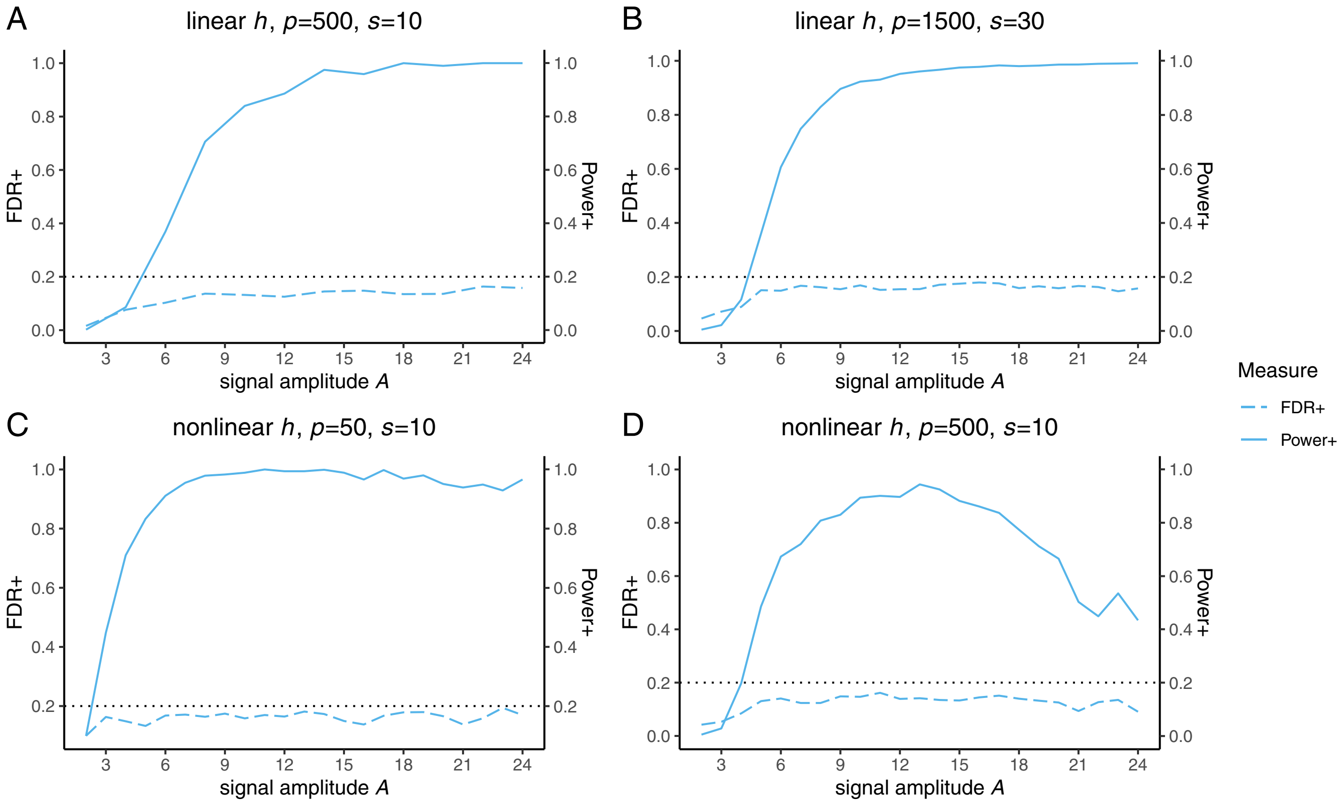

Fig. 5. Simulation results of DeepLINK in settings with the logistic factor model. *A–D* represent different combinations of the link function *h*, the number of features *p*, and the number of true signals *s*. FDR+ and Power+ denote the empirical FDR and power obtained using the knockoff+ threshold. The black dashed lines indicate the target FDR level. Each plot shows the change of FDR+ (solid line) and Power+ (long dashed line) against varying signal amplitude *A*.

https://doi.org/10.1073/pnas.2104683118

and features $\mathbf{x}$, we propose to use the MLP for such modeling purpose. The input of MLP is the augmented data matrix $[\mathbf{X}, \tilde{\mathbf{X}}]$. Instead of directly feeding the augmented feature vector into the MLP, we exploit the idea of DeepPINK developed in ref. 16 and construct a pairwise-connected filter layer with each filter representing a linear combination of one original feature and its knockoff counterpart. The filter layer is then fed to the canonical MLP. The illustrative architecture of DeepPINK is shown in Fig. 2.

To simplify the notation, we use DeepPINK with one hidden layer after the filter layer to discuss the construction of the knockoff statistics. Let $\mathbf{z} = (z_1, \cdots, z_p)^T$ and $\tilde{\mathbf{z}} = (\tilde{z}_1, \cdots, \tilde{z}_p)^T$ be the filter weights (i.e., each filter $F_j = z_j x_j + \tilde{z}_j \tilde{x}_j$) and $\boldsymbol{W}^{(1)} \in \mathbb{R}^{p\times m}$, $\boldsymbol{W}^{(2)} \in \mathbb{R}^{m\times 1}$ be the two weight matrices connecting the filter layer with the output layer, where m is the number of neurons before the output layer. The knockoff statistics are defined as

$$W_j = Z_j^2 - \tilde{Z}_j^2,\ j = 1, \cdots, p, \qquad [13]$$

where $Z_j = z_j w_j$, $\tilde{Z}_j = \tilde{z}_j w_j$, and $\mathbf{w} = (w_1, \cdots, w_p)^T = \boldsymbol{W}^{(1)}\boldsymbol{W}^{(2)}$. This can be easily generalized to cases with more than one hidden layer. Since the weights of neurons are natural measures of their importance, intuitively W_j's defined in Eq. **13** are valid knockoff statistics. Ref. 16 has more detailed discussions on the intuition of W_j in Eq. **13**. Important features can then be selected using the knockoffs inference procedure reviewed previously.

Simulation Studies

We first evaluate the performance of DeepLINK on the simulated datasets. We consider various simulation settings when 1) the factor model is linear or nonlinear, 2) the link function between the response and the features is linear or nonlinear, and 3) the feature dimensionality is low or high. The computational cost of DeepLINK is presented in *SI Appendix*, section 1.

Simulation Designs. We explore three different factor models: the linear factor model (Eq. **14**), the additive quadratic factor model (Eq. **15**), and the logistic factor model (Eq. **16**), where for $i = 1, \dots, n$,

$$\mathbf{x}_i = \mathbf{\Lambda}\mathbf{f}_i + \boldsymbol{\epsilon}_i, \qquad [14]$$

$$\mathbf{x}_i = \mathbf{\Lambda}[\mathbf{f}_i^T,\ (\mathbf{f}_i^2)^T,\ f_{i1}f_{i2},\ f_{i1}f_{i3},\ f_{i2}f_{i3}]^T + \boldsymbol{\epsilon}_i, \qquad [15]$$

$$x_{ij} = \frac{c_j}{1 + \exp([1, \mathbf{f}_i^T]\boldsymbol{\lambda}_j)} + \epsilon_{ij},\ j = 1, \cdots, p. \qquad [16]$$

Here, $\mathbf{f}_i = (f_{i1}, f_{i2}, f_{i3})^T$ is the vector of latent factors, $\mathbf{f}_i^2$ is the shorthand notation for $(f_{i1}^2, f_{i2}^2, f_{i3}^2)^T$, $\mathbf{\Lambda}$ and $\boldsymbol{\lambda}_j$ are the factor loading parameters of appropriate dimensions, and c_j's are some constants. The f_{ij}, c_j, λ_{ij}, and entries of $\mathbf{\Lambda}$ and $\boldsymbol{\epsilon}_i$ are all sampled independently from the standard normal distribution $N(0, 1)$.

The response vector $\mathbf{y} = (y_1, \cdots, y_n)^T$ is simulated from model Eq. **6**. We investigate two different forms of the link function h—the linear design (Eq. **17**) and the nonlinear design (Eq. **18**):

$$h(\mathbf{x}) = \mathbf{x}^T\boldsymbol{\beta}, \qquad [17]$$

$$h(\mathbf{x}) = \sin(\mathbf{x}^T\boldsymbol{\beta})\exp(\mathbf{x}^T\boldsymbol{\beta}). \qquad [18]$$

To simulate the coefficient vector $\boldsymbol{\beta} = (\beta_1, \cdots, \beta_p)^T$, we first randomly choose s true signal locations and then set the β_j at

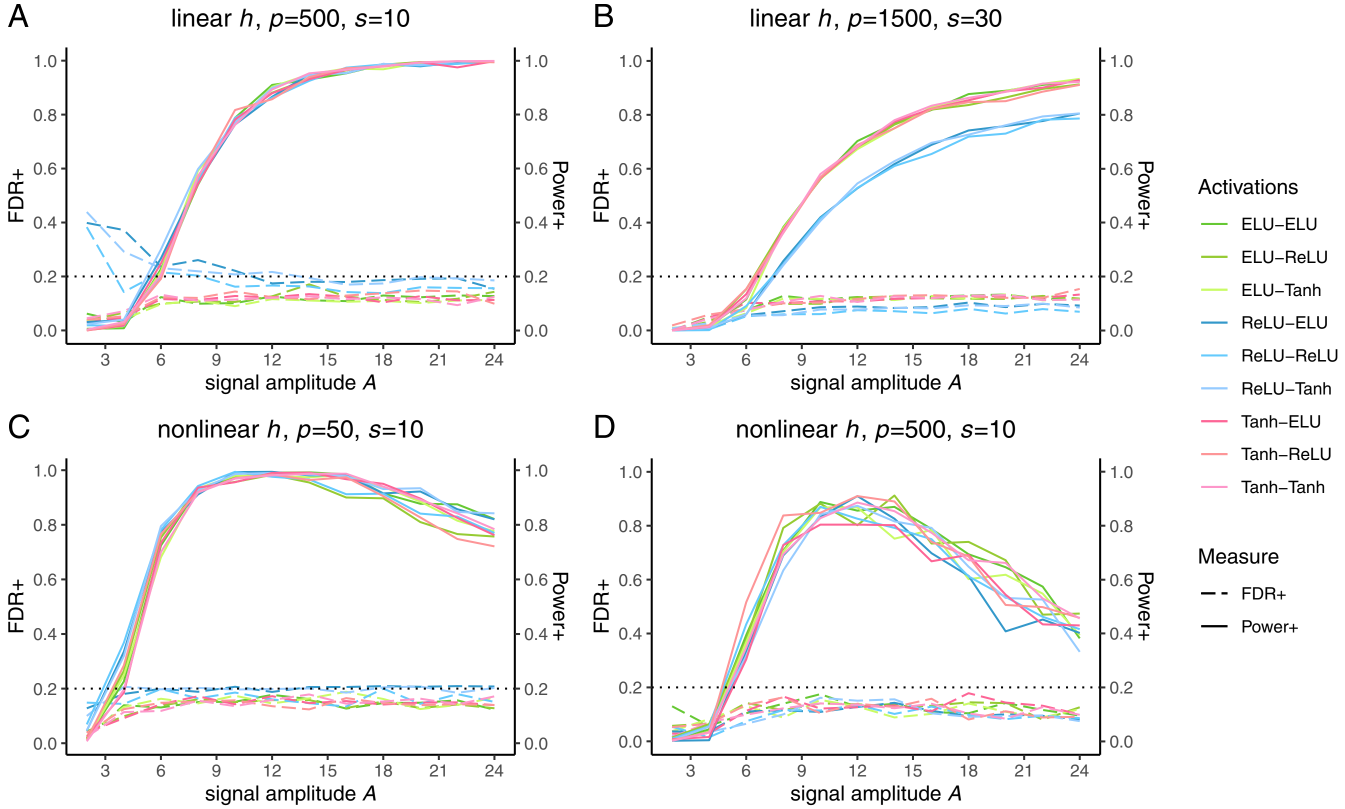

Fig. 6. Linear factor model simulation results of DeepLINK using different activation functions. *A–D* represent different combinations of the link function *h*, the number of features *p*, and the number of true signals *s*. FDR+ and Power+ denote the empirical FDR and power obtained using the knockoff+ threshold. The black dashed lines indicate the target FDR level. Each plot shows the change of FDR+ (solid lines) and Power+ (long dashed lines) against varying signal amplitude *A* for different activation functions (ELU, ReLU, and Tanh) used in autoencoder and DeepPINK (e.g., ReLU-ELU represents using ReLU in autoencoder and ELU in DeepPINK).

each location to be A or $-A$ with equal probability, where A is some positive value that varies in our simulation studies. The remaining $p-s$ components of $\boldsymbol{\beta}$ are set to zero. It is seen that when the link function h is linear, A measures the signal strength with a larger value corresponding to a stronger signal. When h is nonlinear, however, the signal strength may no longer be a monotone increasing function of A. The discussions in *SI Appendix*, section 2 have an example illustrating this. In fact, to the best of our knowledge, there lacks a widely adopted measure for the signal strength in the nonlinear model settings. Model errors ε_is are also sampled independently from the standard normal distribution $N(0,1)$.

Parameter Settings. For all the simulation studies, the target FDR q is set to 0.2, and the sample size n is set to 1,000. For the linear link function setting, we explore two different feature dimensionalities $p=500, 1{,}500$ with true signal size s set to 10 and 30, respectively. For the nonlinear link function setting, p is set to 50 and 500, and s is fixed at 10. We vary the value of A to investigate its impact on the performance of DeepLINK.

Neural Network Settings. We next provide the details on the neural network architectures. We train the autoencoder network using the Adam algorithm with the mean squared error (MSE) as the loss function. For the linear factor model, the number of neurons in the autoencoder's bottleneck layer is estimated by the PC_{p1} algorithm proposed in ref. 43. It is worth noting that PC_{p1} is designed for linear factor models. For the nonlinear factor models, we set it to the true number of factors $r=3$. We conduct a robustness study of DeepLINK to the misspecification of r in *SI Appendix*, section 3. We remark that r can be tuned by the cross-validation in real applications. For DeepPINK used in the feature selection step, we use MSE as the loss function coupled with the L_1 regularization when the link function h is linear. When the link function is nonlinear, we change the loss function to the mean squared logarithmic error (MSLE) because MSE may cause explosive gradients for large response values. In fact, MSLE also works well with other nonlinear link functions (*SI Appendix*, section 4). For a general guidance, we suggest using MSE first and switching to MSLE when the gradients become too large during the model training. For both linear and nonlinear link functions, we use the Adam optimizer to train the network. For both autoencoder and DeepPINK networks, we recommend to use the exponential linear unit (ELU) as the activation function according to our experience gained from empirical studies. Our numerical study also suggests that the learning rate of Adam and the coefficient of L_1 regularization need to be tuned for the best performance of our method. The neural network settings are summarized in Table 1.

Simulation Results. We investigate the performance of DeepLINK in the simulation study with different combinations of factor models, link functions, and dimensionalities. For each setting, we apply DeepLINK to 100 independently simulated datasets and calculate the average FDP and TDP as the empirical FDR and power, respectively. The knockoff+ threshold is used in our numerical studies because it controls the exact FDR.

Simulation results with the linear factor model. We compare DeepLINK with IPAD reviewed in the Introduction using the simulated data in the linear factor model setting as specified in Eq. **14**. Both methods successfully control the FDR under the target level 0.2. In terms of power, IPAD slightly

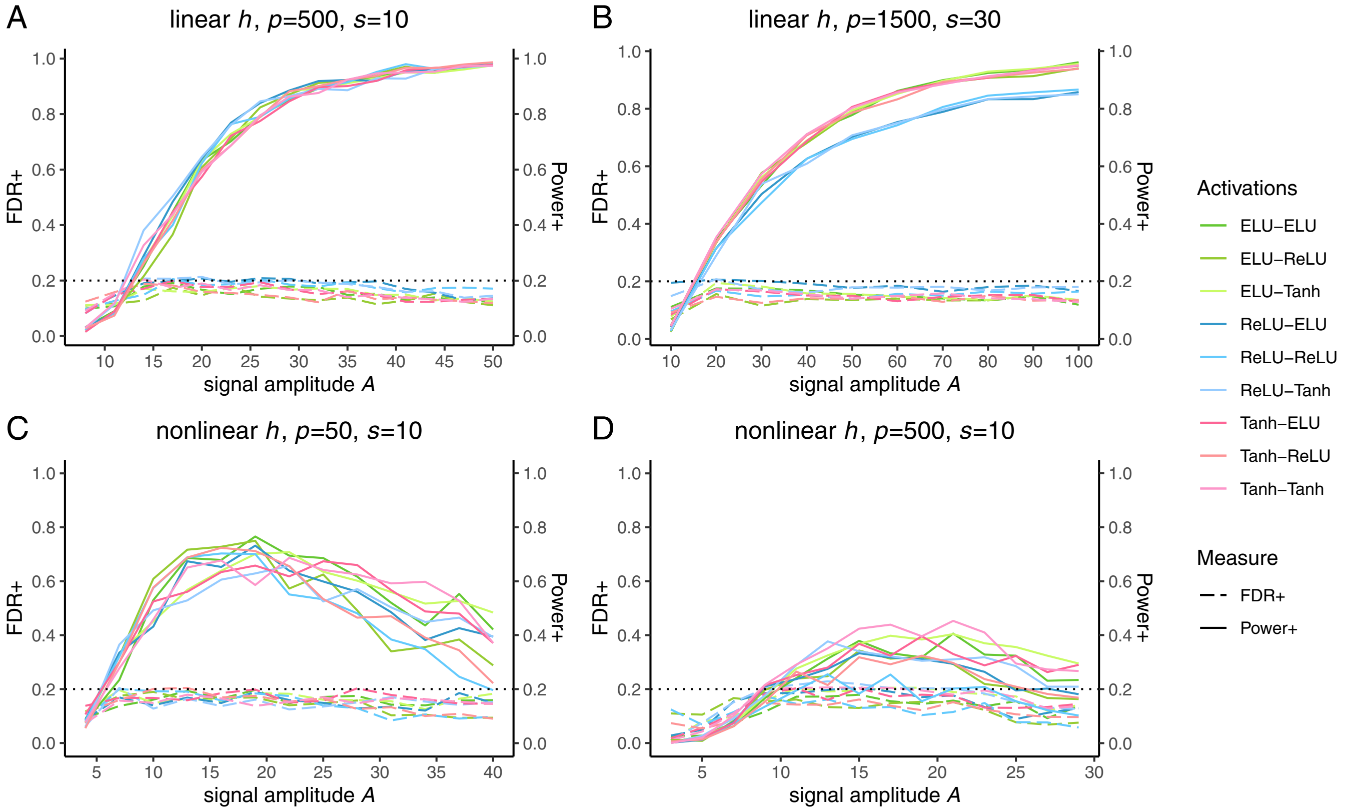

Fig. 7. Additive quadratic factor model simulation results of DeepLINK using different activation functions. *A–D* represent different combinations of the link function *h*, the number of features *p*, and the number of true signals *s*. FDR+ and Power+ denote the empirical FDR and power obtained using the knockoff+ threshold. The black dashed lines indicate the target FDR level. Each plot shows the change of FDR+ (solid lines) and Power+ (long dashed lines) against varying signal amplitude *A* for different activation functions (ELU, ReLU, and Tanh) used in autoencoder and DeepPINK (e.g., ReLU-ELU represents using ReLU in autoencoder and ELU in DeepPINK).

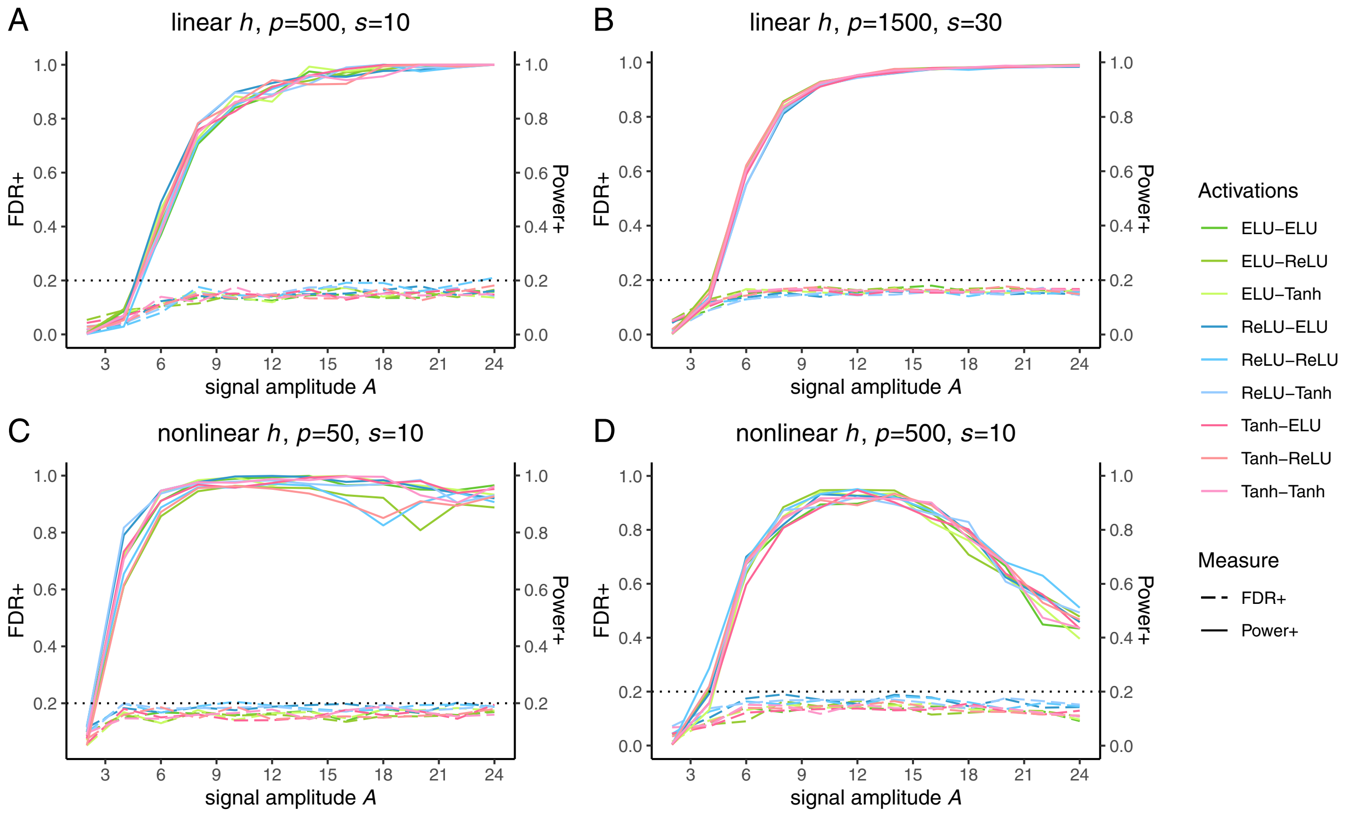

Fig. 8. Logistic factor model simulation results of DeepLINK using different activation functions. *A–D* represent different combinations of the link function *h*, the number of features *p*, and the number of true signals *s*. FDR+ and Power+ denote the empirical FDR and power obtained using the knockoff+ threshold. The black dashed lines indicate the target FDR level. Each plot shows the change of FDR+ (solid lines) and Power+ (long dashed lines) against varying signal amplitude *A* for different activation functions (ELU, ReLU, and Tanh) used in autoencoder and DeepPINK (e.g., ReLU-ELU represents using ReLU in autoencoder and ELU in DeepPINK).

outperforms DeepLINK in settings with the linear link function (Fig. 3 *A* and *B*). This is reasonable because IPAD was proposed under the assumption of the linear factor model and linear link function and makes full use of these parametric model structures, while DeepLINK makes no use of these model structures at all. For the nonlinear link function (Fig. 3 *C* and *D*), however, the power of IPAD drops significantly, while DeepLINK still maintains decently high power. It is also interesting to observe that the power of DeepLINK first increases sharply to the peak and then decreases slightly as A increases, which can be explained by the fact that A no longer serves as a good measure of the signal strength. These results demonstrate the versatility of DeepLINK. The capability of DeepLINK to tackle complicated nonlinear link functions makes it more useful in real applications since it is more robust to possible model misspecification.

Another interesting observation is that our simulation produces highly correlated features. To study DeepLINK's ability to disentangle important features from their highly correlated noise features, we conducted additional analyses in *SI Appendix*, section 5.

Simulation results with the nonlinear factor model. We now consider nonlinear factor models Eqs. **15** and **16**. We will drop IPAD from the comparison because IPAD was proposed under the assumption of linear factor model and is not expected to perform well when the model is severely misspecified.* For the additive quadratic factor model in Eq. **15**, FDR is perfectly controlled in all settings. Meanwhile, high power is achieved with reasonably large A in the two settings with linear link function (Fig. 4 *A* and *B*). However, the two settings with nonlinear link functions are very challenging, and the power of DeepLINK is significantly lower (Fig. 4 *C* and *D*). For the logistic factor model Eq. **16**, DeepLINK controls the FDR and can achieve power close to one with a wide range of values for A in each setting (Fig. 5). The success of FDR control by DeepLINK in nonlinear factor model settings provides evidence that the autoencoder network can well capture the nonlinear factor structure and thus, generates valid knockoffs data matrices. Similar to the linear factor model setting, we again observe an inverted U-shaped curve of the power when the link function is nonlinear, which can be explained by the same reason as before.

Robustness of DeepLINK to different activation functions. We explore the effects of different activation functions used in the autoencoder and DeepPINK networks on the performance of DeepLINK (Figs. 6–8). In general, DeepLINK is robust to different combinations of activation functions in terms of both FDR

Table 2. Mean and SE (in parentheses) of the misclassification error rates for the microbiome dataset

	Training	Test
$d = 20$	0.172 (0.003)	0.319 (0.008)
$d = 30$	0.104 (0.004)	0.306 (0.007)
$d = 40$	0.019 (0.003)	0.328 (0.009)
$d = 50$	0.008 (0.002)	0.319 (0.008)
$d = 100$	0.000 (0.000)	0.385 (0.012)

*The performance of IPAD is already very poor when the link function *h* alone takes a nonlinear form, as shown in Fig. 3 *C* and *D*.

Table 3. Top 20 most selected microbial species when $d = 30$ for the microbiome dataset

Species	Frequency
D. pneumosintes	99
Eikenella corrodens	96
Staphylococcus haemolyticus	75
Intestinimonas butyriciproducens	75
B. fragilis	70
Latilactobacillus sakei	58
Clostridium bornimense	58
P. micra	51
A. muciniphila	49
Gemella sp. oral taxon 928	48
Clostridium chauvoei	45
Corynebacterium sp. NML98-0116	43
Prevotella intermedia	34
Streptococcus sp. A12	31
Ndongobacter massiliensis	30
Ruminococcus bicirculans	29
Lactococcus garvieae	28
Fusobacterium varium	26
Anaerococcus mediterraneensis	26
Desulfovibrio fairfieldensis	23

control and power. The only exception is using the rectified linear unit (ReLU) activation in the autoencoder network. In the linear factor model setting with linear link function h and low feature dimensionality, the autoencoder with ReLU activation fails to control the FDR when the signal amplitude is small (Fig. 6*A*). We also observe that the autoencoder with ReLU activation has slightly inflated FDR in some other settings (Figs. 7 *A* and *D* and 8*A*). In the linear and additive quadratic factor model settings with linear link function h and large feature dimensionality (Figs. 6*B* and 7*B*), ReLU yields lower power than other activation functions when used in the autoencoder network. We thus recommend against using the ReLU activation in the autoencoder network for the DeepLINK applications.

Real Data Applications

We further apply DeepLINK to three real data applications. All predictors in the three datasets below were standardized to unit variance before the analysis. In all real data applications, the error distribution $f_{\boldsymbol{\theta}}$ was fitted assuming Gaussian distribution. The robustness of DeepLINK with respect to misspecified error distribution is investigated in *SI Appendix*, section 6. We also compare the performance of DeepLINK with that of random forests (44, 45) in *SI Appendix*, section 7.

Application to a Microbiome Dataset. The microbiome dataset is publicly available in a colorectal cancer (CRC)–related metagenomic study in Zeller et al. (46). The dataset contains the whole genome–sequenced (WGS) DNAs from stool samples of 184 individuals (91 CRC patients and 93 healthy controls). We aligned the DNA sequences against the National Center for Biotechnology Information (NCBI) microbial reference genome database and constructed an abundance matrix according to the alignment results. The matrix consisted of 184 rows and 434 columns, with each entry representing the abundance of a microbial species in the corresponding sample. We randomly split the dataset into the training set (80%) and the test set (20%) and implemented the DeepLINK on the training part. The trained model was then applied to the test data, and the classification error rate was calculated. The random splitting procedure was repeated 100 times. However, the mean misclassification error on the test data was consistently around 0.5 under various parameter settings of DeepLINK, suggesting that the simple application of DeepLINK could fail. We also tried some other popular classification methods such as the Lasso and simple deep neural network without the special architecture as in DeepLINK, all of which gave us error rates between 0.4 and 0.5, similar to the random guessing.

Consequently, we performed a variable screening step first and then applied the DeepLINK method on the screened dataset. Considering the relatively small sample size and to reduce the chance of including noise confounders, we identified an independent microbiome dataset for screening. This independent dataset was also publicly available and was collected for CRC–microbiome association analysis (47). It contained 128 WGS DNA samples with 74 CRC patients and 54 controls. Since these two microbiome datasets had different numbers of features, we constrained ourselves to the 274 common features in our analysis. There are multiple options for the screening step (2, 48, 49). We adopted one of the state-of-the-art methods, which was based on the distance correlation, and ranked these 274 variables by the values of the asymptotic test statistics (50, 51). We randomly split the Zeller et al. (46) CRC microbiome data into the training and test sets at the ratio of 80 to 20%. The justification for the training/testing set split ratio is given in *SI Appendix*, section 8. Then, we trained the DeepLINK model using the top-ranked variables with the training data. To evaluate the impact of the number of retained variables after the screening step, denoted as d, we examined multiple values of d. Finally, we applied the trained DeepLINK model onto the training and test datasets and calculated the corresponding classification error rates. The whole process was repeated 100 times. We set the number of neurons in the bottleneck layer of the autoencoder to three. We chose the other model parameters by cross-validation. The MLP in DeepPINK had only one hidden layer with d neurons. The dropout rate was 0.4, while L_1- and L_2-regularization weights were both 0.001. The mean and SE of the misclassification error on the training and test data are given in Table 2. We can see that the mean test error was the lowest when 30 variables were retained after the screening step. However, as d increased to as large as 100, the mean test error became relatively high (0.385), indicating that when DeepLINK lost the help of the screening step in eliminating noise variables, its performance could be compromised.

The top 20 most selected microbial species along with their selection frequencies by DeepLINK coupled with screening are presented in Table 3. We only present the results for $d = 30$ when the mean classification error rate was the lowest. Many of these selected species were reported to have important associations with CRC in the previous literature. For example, *Parvimonas micra* and *Akkermansia muciniphila* were among the four-bacteria biomarker panel of CRC identified by Osman et al. (52). In addition, *P. micra*'s enrichment in CRC was demonstrated in a number of previous studies (47, 53–55), and Purcell et al. (56) also reported its enrichment in one of the CRC

Table 4. Mean and SE (in parentheses) of the misclassification error rates for the murine scRNA-seq dataset

	Training	Test
$d = 20$	0.000 (0.000)	0.021 (0.003)
$d = 30$	0.000 (0.000)	0.018 (0.002)
$d = 40$	0.000 (0.000)	0.012 (0.002)
$d = 50$	0.000 (0.000)	0.014 (0.002)
$d = 100$	0.000 (0.000)	0.016 (0.002)
$d = 200$	0.000 (0.000)	0.010 (0.001)
$d = 300$	0.000 (0.000)	0.013 (0.002)
$d = 400$	0.000 (0.000)	0.012 (0.002)
$d = 500$	0.000 (0.000)	0.015 (0.002)

Table 5. Top 20 most selected genes when $d = 200$ for the murine scRNA-seq dataset

Gene	Frequency	Gene	Frequency
Sqstm1	73	*Gm26825*	61
Cdkn1a	66	*Hsp90aa1*	60
Sdc4	65	*Tnfaip2*	60
Abcg1	64	*Clec4e*	58
Rab31	64	*Gpx1*	58
Gm28875	64	*Sod2*	57
Gmnn	63	*Srsf5*	55
Angpt2	63	*Fas*	51
Ehd4	61	*Get1*	50
Dnaja1	61	*Hsp90ab1*	50

subtypes. Other important CRC-related species that were also reported in previous studies include *Dialister pneumosintes* (57) and *Bacteroides fragilis* (58–61).

Application to a Murine Single-Cell RNA-Sequencing Dataset. The murine single-cell RNA-sequencing (scRNA-seq) dataset is publicly available from Lane et al. (62), aiming to investigate the effect of lipopolysaccharides (LPS)-stimulated nuclear factor-κB (NF-κB) on gene expression. We first preprocessed the data following the suggestions in ref. 63. We filtered out cells either with mapping rate below 20% or with nonzero expression proportion below 5%. We also filtered out genes expressed in less than 5% of total cells. The preprocessed data matrix contained the expression, in the form of transcripts per million (TPM), of 13,777 genes from 570 cells. We were interested in differential gene expression between cells with two conditions: unstimulated (202 cells) and stimulated with LPS after 150 min (368 cells). Due to the high dimensionality, it was computationally infeasible to implement the DeepLINK on this dataset directly, even with powerful servers. The success of screening in the previous microbiome example motivated us to apply a screening step first to reduce the dimensionality. Since this dataset had a relatively larger sample size than the microbiome dataset, we randomly split the dataset into three parts for screening (50%), training (40%), and test (10%), instead of using an independent dataset for screening. We used the same model architecture and parameters tuned from the previous microbiome analysis. The mean and SE of the misclassification error over 100 repetitions on the training and test sets, respectively, for this scRNA-seq dataset are provided in Table 4. We observe that the mean misclassification error on the test data can get as low as 0.010 when $d = 200$.

We further looked at the top 20 most selected genes by DeepLINK equipped with screening for $d = 200$ as presented in Table 5. Many of the selected genes were also reported as significant features in the original study (62) including Sqstm1, Sdc4, Abcg1, Rab31, Gmnn, Angpt2, Hsp90aa1, Tnfaip2, Clec4e, Gpx1, Sod2, and Fas. Gene Ontology (GO) analysis with domain Biological Process (BP) indicates that the up-regulation of genes in LPS-stimulated cells is related to NF-κB signaling (Sqstm1) and LPS response (Sod2). Also, Hsp90aa1 can bind LPS and mediate LPS-induced inflammatory response according to Uniprot (64), which may be related to its up-regulation in LPS-stimulated cells.

Application to a Human Single-Cell RNA-Sequencing Dataset. Another scRNA-seq dataset that we investigated is from a human glioblastoma study led by Darmanis et al. (65). We were interested in differential gene expression between Neoplastic cells in the tumor core and the surrounding periphery. We used the same criteria as in the murine scRNA-seq study to preprocess the data, which resulted in a dataset with TPMs of 23,257 genes from 632 cells (580 in the tumor core and 52 in the periphery). Again, due to the high dimensionality, we first conducted dimensionality reduction using the distance correlation screening and then applied DeepLINK. The model architecture and parameters were the same as those in the last two real data studies. We repeated the experiment 100 times and present the mean and SE of the misclassification error on the training and test data, respectively, in Table 6. We see that the mean misclassification errors on the test data achieve the smallest value when $d = 200$ and then become more or less stable.

The top 20 most selected genes by DeepLINK equipped with screening for $d = 200$ are shown in Table 7. We next examined the biological meaning of these selected genes. As pointed out in the original study (65), down-regulation of genes like ATP1A2 and PRODH in the periphery might be related to their functions in the interstitial matrix invasion. We also observed that HIF3A was down-regulated in the tumor core, which was probably associated with the hypoxia in core. Previous study also demonstrated that HIF3A was a dominant-negative regulator of HIF-1 and was thus down-regulated in a hypoxic environment (66). GO analysis with domain BP indicates that some genes up-regulated in periphery have functions related to cell migration from periphery to core. For instance, HES6 has GO term nervous system development, which is highly relevant to tumor cell migration. IGSF21 and CNTN1 have GO term cell–cell adhesion, which is a central part in cell migration. ALDOC has GO term glycolytic process, which produces a small amount of adenosine triphosphate (ATP) and may help the cell migration as an energy provider. SERPINE2 has GO term regulation of cell migration. Also, genes such as SPARCL1, NPL, and ST6GALNAC3 are involved in various metabolic processes.

Discussion

In this paper, we have developed a high-dimensional inference framework via knockoffs, DeepLINK, to enhance the interpretability and reproducibility of deep learning models. DeepLINK generates the knockoff variables under the possibly nonlinear factor model assumption using an autoencoder network and then fits the regression/classification model using the DeepPINK network. We have used various simulated datasets to numerically demonstrate that DeepLINK can achieve successful FDR control with attractive power in selecting features that are truly important for the response of interest. We have also showcased the practical utility and performance of DeepLINK on three real data applications.

When comparing the prediction performance of DeepLINK with random forests in *SI Appendix*, section 7, we noticed that random forests can outperform DeepLINK in terms of prediction for the microbiome dataset. This is likely caused by the distinctive prediction power of MLP and random forests. We remark that the MLP in the second step of DeepLINK can be replaced with random forests if one suspects that the latter can outperform in prediction. We also emphasize that the main

Table 6. Mean and SE (in parentheses) of the misclassification error for the human scRNA-seq dataset

	Training	Test
$d = 20$	0.006 (0.001)	0.072 (0.003)
$d = 30$	0.002 (0.000)	0.068 (0.003)
$d = 40$	0.001 (0.000)	0.064 (0.003)
$d = 50$	0.001 (0.000)	0.060 (0.003)
$d = 100$	0.001 (0.000)	0.059 (0.003)
$d = 200$	0.000 (0.000)	0.046 (0.003)
$d = 300$	0.000 (0.000)	0.056 (0.003)
$d = 400$	0.000 (0.000)	0.049 (0.003)
$d = 500$	0.000 (0.000)	0.050 (0.003)

Table 7. Top 20 most selected genes when $d = 200$ for the human scRNA-seq dataset

Gene	Frequency	Gene	Frequency
ATP1A2	73	*ANKRD20A9P*	45
PRODH	73	*HIF3A*	44
HES6	62	*NPL*	44
IGSF21	57	*AC131097.1*	44
PPM1K	56	*FAM240C*	41
ALDOC	55	*ST6GALNAC3*	40
SPARCL1	55	*MMP28*	40
SERPINE2	52	*CNTN1*	39
RNPC3	52	*MTRNR2L1*	39
LOC102724788	47	*EFHD1*	38

purpose of DeepLINK is feature selection with controlled error rate, and to achieve the goal of FDR control in feature selection, the prediction power may be slightly compromised in some applications.

There are five potential directions for future investigations. First, in the real data applications, we consider binary outcomes. DeepLINK can be easily extended to the case of multiple classes if we replace the loss function in the second step of binary cross-entropy with multiclass cross-entropy. Second, the knockoff variable-generating process of DeepLINK simulates the idiosyncratic matrix **E** outside of the autoencoder network with nondeep learning techniques. Designing a new deep neural network, which can automate the knockoff variable-generating process, may increase its efficiency and accuracy. Third, we currently have two separate networks in DeepLINK: the knockoff variable-generating network of autoencoder and the model fitting and inference network of DeepPINK. We would like to integrate them into one single network for a joint optimization so that the whole process can be fully automated. Such a feature can make DeepLINK even more user friendly. Fourth, heterogeneity in the samples is a practically important issue. It is possible that the samples consist of multiple subpopulations and that they have different true features. It is likely that DeepLINK can be extended to accommodate the heterogeneity. The key is to construct valid knockoff variables reflecting the subpopulation information. One naive method is to construct knockoff variables for each subpopulation and then combine them appropriately to form valid knockoff variables for the overall population. If this can be achieved, the second step of feature selection using MLP can be applied without modification. Finally, we would like to provide theoretical justifications on DeepLINK in terms of both FDR control and power. This can in turn guide the training of the underlying networks and further improve the interpretation of our deep learning inference method.

Data Availability. Software data have been deposited in GitHub (https://github.com/zifanzhu/DeepLINK). Preprocessed data matrices for the four publicly available data sets can be downloaded with the corresponding link: Zeller microbiome (67), Yu microbiome (68), murine scRNA-seq (69), and human scRNA-seq (70).

ACKNOWLEDGMENTS. This work was supported by NIH Grants R01GM120624 and 1R01GM131407 and NSF Grant DMS-1953356. Z.Z. was supported by the Viterbi Fellowship.

1. J. Fan, J. Lv, A selective overview of variable selection in high dimensional feature space. *Stat. Sin.* **20**, 101–148 (2010).
2. J. Fan, Y. Fan, High-dimensional classification using features annealed independence rules. *Ann. Stat.* **36**, 2605–2637 (2008).
3. Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
4. Y. Benjamini, Discovering the false discovery rate. *J. R. Stat. Soc. B* **72**, 405–416 (2010).
5. Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* **29**, 1165–1188 (2001).
6. B. Efron, R. Tibshirani, J. D. Storey, V. Tusher, Empirical Bayes analysis of a microarray experiment. *J. Am. Stat. Assoc.* **96**, 1151–1160 (2001).
7. J. D. Storey, A direct approach to false discovery rates. *J. R. Stat. Soc. Series B Stat. Methodol.* **64**, 479–498 (2002).
8. Y. Benjamini, A. M. Krieger, D. Yekutieli, Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* **93**, 491–507 (2006).
9. C. R. Genovese, K. Roeder, L. Wasserman, False discovery control with p-value weighting. *Biometrika* **93**, 509–524 (2006).
10. J. G. Scott, R. C. Kelly, M. A. Smith, P. Zhou, R. E. Kass, False discovery rate regression: An application to neural synchrony detection in primary visual cortex. *J. Am. Stat. Assoc.* **110**, 459–471 (2015).
11. N. Ignatiadis, B. Klaus, J. B. Zaugg, W. Huber, Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nat. Methods* **13**, 577–580 (2016).
12. M. Stephens, False discovery rates: A new deal. *Biostatistics* **18**, 275–294 (2017).
13. L. Lei, W. Fithian, Adapt: An interactive procedure for multiple testing with side information. *J. R. Stat. Soc. Series B Stat. Methodol.* **80**, 649–679 (2018).
14. A. Li, R. F. Barber, Multiple testing with the structure-adaptive Benjamini-Hochberg algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.* **81**, 45–74 (2019).
15. E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: ‘Model-X’ knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. B* **80**, 551–577 (2018).
16. Y. Lu, Y. Fan, J. Lv, W. S. Noble, “DeepPINK: Reproducible feature selection in deep neural networks” in *Advances in Neural Information Processing Systems*, S. Bengio *et al.*, Eds. (Advances in Neural Information Processing Systems, 2018), pp. 8676–8686.
17. Y. Uematsu, Y. Fan, K. Chen, J. Lv, W. Lin, SOFAR: Large-scale association network learning. *IEEE Trans. Inf. Theory* **65**, 4924–4939 (2019).
18. Z. Zheng, J. Lv, W. Lin, Nonsparse learning with latent variables. *Oper. Res.* **69**, 346–359 (2021).
19. C. Friguet, M. Kloareg, D. Causeur, A factor model approach to multiple testing under dependence. *J. Am. Stat. Assoc.* **104**, 1406–1415 (2009).
20. Y. Shen, R. Jin, “Learning personal+ social latent factor model for social recommendation” in *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Q. Yang, D. Agarwal, Eds. (Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2012), pp. 1303–1311.
21. R. Jenatton, N. Le Roux, A. Bordes, G. Obozinski, “A latent factor model for highly multi-relational data” in *Advances in Neural Information Processing Systems (NIPS 2012)*, F. Pereira, C. J. C. Burges, L. Bottou, K. Q. Weinberger, Eds. (Advances in Neural Information Processing Systems, 2012), vol. 25, pp. 3176–3184.
22. R. Argelaguet *et al.*, Multi-Omics Factor Analysis—A framework for unsupervised integration of multi-omics data sets. *Mol. Syst. Biol.* **14**, e8124 (2018).
23. E. Frichot, S. D. Schoville, G. Bouchard, O. François, Testing for associations between loci and environmental gradients using latent factor mixed models. *Mol. Biol. Evol.* **30**, 1687–1699 (2013).
24. Y. Blum, G. Le Mignon, S. Lagarrigue, D. Causeur, A factor model to analyze heterogeneity in gene expression. *BMC Bioinformatics* **11**, 368 (2010).
25. J. Fan, Y. Fan, J. Lv, High dimensional covariance matrix estimation using a factor model. *J. Econom.* **147**, 186–197 (2008).
26. J. T. Scott, Factor analysis and regression. *Econometrica* **34**, 552–562 (1966).
27. J. T. Scott, Factor analysis regression revisited. *Econometrica* **37**, 719 (1969).
28. F. X. Diebold, G. D. Rudebusch, S. B. Aruoba, The macroeconomy and the yield curve: A dynamic latent factor approach. *J. Econom.* **131**, 309–338 (2006).
29. A. Uddin, D. Yu, Latent factor model for asset pricing. *J. Behav. Exp. Finance* **27**, 100353 (2020).
30. Y. Fan, J. Lv, M. Sharifvaghefi, Y. Uematsu, IPAD: Stable interpretable forecasting with knockoffs inference. *J. Am. Stat. Assoc.* **115**, 1822–1834 (2020).
31. J. Pearl, “Markov and Bayesian networks: Two graphical representations of probabilistic knowledge” in *Probabilistic Reasoning in Intelligent Systems*, J. Pearl, Ed. (Morgan Kaufmann, San Francisco, CA, 1988), pp. 77–141.
32. G. Hommel, T. Hoffmann, *Controlled Uncertainty in Multiple HypothesenprüfungIMultiple Hypotheses Testing* (Springer, 1988), pp. 154–161.
33. E. L. Lehmann, J. P. Romano, *Generalizations of the Familywise Error Rate in Selected Works of EL Lehmann* (Springer, 2012), pp. 719–735.
34. Y. Fan, E. Demirkaya, J. Lv, Nonuniformity of p-values can occur early in diverging dimensions. *J. Mach. Learn. Res.* **20**, 1–33 (2019).
35. R. F. Barber, E. J. Candès, Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).
36. Y. Fan, E. Demirkaya, G. Li, J. Lv, RANK: Large-scale inference with graphical nonlinear knockoffs. *J. Am. Stat. Assoc.* **115**, 362–379 (2020).
37. S. Bates, E. Candès, L. Janson, W. Wang, Metropolized knockoff sampling. *J. Am. Stat. Assoc.*, (2020).
38. D. Huang, L. Janson, Relaxing the assumptions of knockoffs by conditioning. *Ann. Stat.* **48**, 3021–3042 (2020).
39. M. Sesia, C. Sabatti, E. J. Candès, Gene hunting with hidden Markov model knockoffs. *Biometrika* **106**, 1–18 (2019).
40. X. Bai, J. Ren, Y. Fan, F. Sun, KIMI: Knockoff inference for motif identification from molecular sequences with controlled false discovery rate. *Bioinformatics* **37**, 759–766 (2021).
41. I. Jolliffe, *Principal Component Analysis* (Springer Verlag, New York, NY, 2002).

42. Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* **35**, 1798–1828 (2013).
43. J. Bai, S. Ng, Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221 (2002).
44. L. Breiman, Random forests. *Mach. Learn.* **45**, 5–32 (2001).
45. C. M. Chi, P. Vossler, Y. Fan, J. Lv, Asymptotic properties of high-dimensional random forests. arXiv [Preprint] (2020). https://arxiv.org/abs/2004.13953 (Accessed 29 April 2020).
46. G. Zeller *et al*., Potential of fecal microbiota for early-stage detection of colorectal cancer. *Mol. Syst. Biol.* **10**, 766 (2014).
47. J. Yu *et al*., Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
48. J. Fan, J. Lv, Sure independence screening for ultrahigh dimensional feature space (with discussion). *J. R. Stat. Soc. B* **70**, 849–911 (2008).
49. J. Fan, J. Lv, Sure independence screening (invited review article). *Wiley StatsRef: Statistics Reference Online* (2018). https://par.nsf.gov/biblio/10091881. Accessed 1 June 2018.
50. G. J. Székely, M. L. Rizzo, N. K. Bakirov, Measuring and testing dependence by correlation of distances. *Ann. Stat.* **35**, 2769–2794 (2007).
51. L. Gao, Y. Fan, J. Lv, Q. Shao, Asymptotic distributions of high-dimensional distance correlation inference. *Ann. Stat.* (2021).
52. M. A. Osman *et al*., *Parvimonas micra*, *Peptostreptococcus stomatis*, *Fusobacterium nucleatum* and *Akkermansia muciniphila* as a four-bacteria biomarker panel of colorectal cancer. *Sci. Rep.* **11**, 2925 (2021).
53. T. Löwenmark *et al*., Parvimonas micra as a putative non-invasive faecal biomarker for colorectal cancer. *Sci. Rep.* **10**, 15250 (2020).
54. Z. Dai *et al*., Multi-cohort analysis of colorectal cancer metagenome identified altered bacteria across populations and universal bacterial markers. *Microbiome* **6**, 70 (2018).
55. J. L. Drewes *et al*., High-resolution bacterial 16S rRNA gene profile meta-analysis and biofilm status reveal common colorectal cancer consortia. *NPJ Biofilms Microbiomes* **3**, 34 (2017).
56. R. V. Purcell, M. Visnovska, P. J. Biggs, S. Schmeier, F. A. Frizelle, Distinct gut microbiome patterns associate with consensus molecular subtypes of colorectal cancer. *Sci. Rep.* **7**, 11590 (2017).
57. L. Ai *et al*., Systematic evaluation of supervised classifiers for fecal microbiota-based prediction of colorectal cancer. *Oncotarget* **8**, 9546–9556 (2017).
58. N. U. Toprak *et al*., A possible role of *Bacteroides fragilis* enterotoxin in the aetiology of colorectal cancer. *Clin. Microbiol. Infect.* **12**, 782–786 (2006).
59. A. Boleij *et al*., The *Bacteroides fragilis* toxin gene is prevalent in the colon mucosa of colorectal cancer patients. *Clin. Infect. Dis.* **60**, 208–215 (2015).
60. K. S. Viljoen, A. Dakshinamurthy, P. Goldberg, J. M. Blackburn, Quantitative profiling of colorectal cancer-associated bacteria reveals associations between fusobacterium spp., enterotoxigenic *Bacteroides fragilis* (ETBF) and clinicopathological features of colorectal cancer. *PLoS One* **10**, e0119462 (2015).
61. F. Haghi, E. Goli, B. Mirzaei, H. Zeighami, The association between fecal enterotoxigenic *B. fragilis* with colorectal cancer. *BMC Cancer* **19**, 879 (2019).
62. K. Lane *et al*., Measuring signaling and RNA-seq in the same cell links gene expression to dynamic patterns of nf-κb activation. *Cell Syst.* **4**, 458–469.e5 (2017).
63. K. Korthauer *et al*., A practical guide to methods controlling false discoveries in computational biology. *Genome Biol.* **20**, 118 (2019).
64. UniProt Consortium, UniProt: A hub for protein information. *Nucleic Acids Res.* **43**, D204–D212 (2015).
65. S. Darmanis *et al*., Single-cell RNA-seq analysis of infiltrating neoplastic cells at the migrating front of human glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).
66. M. A. Maynard *et al*., Human HIF-3alpha4 is a dominant-negative regulator of HIF-1 and is down-regulated in renal cell carcinoma. *FASEB J.* **19**, 1396–1406 (2005).
67. Z. Zhu, Data from "Abundance matrix for the WGS microbiome data set from Zeller et al." GitHub. https://github.com/zifanzhu/DeepLINK/tree/main/Real_data_analyses/human_microbiome/data/microbiome_data_common.csv. Accessed 21 June 2021.
68. Z. Zhu, Data from "Abundance matrix for the WGS microbiome data set from Yu et al." GitHub. https://github.com/zifanzhu/DeepLINK/tree/main/Real_data_analyses/human_microbiome/data/yu_CRC_common.csv. Accessed 21 June 2021.
69. Z. Zhu, Data from "Expression matrix for the murine scRNA-seq data set from Lane et al." GitHub. https://github.com/zifanzhu/DeepLINK/tree/main/Real_data_analyses/murine_sc_RNAseq/data/rna1.csv. Accessed 21 June 2021.
70. Z. Zhu, Data from "Expression matrix for the human scRNA-seq data set from Darmanis et al." GitHub. https://github.com/zifanzhu/DeepLINK/tree/main/Real_data_analyses/human_sc_RNAseq/data/rna2.csv. Accessed 21 June 2021.